TAMIL THESAURUS TO TAMIL WORDNET

Rajendran Sankaravelayuthan
Amrita Vishwa Vidyapeetham , Coimbatore
rajushush@gmail.com

Abstract

Tamil Onto-thesaurus is an outcome of a very long research activity that went on in the field of lexical semantics of Tamil vocabulary. It went through several stages before being culminated into Tamil onto-thesaurus. It depicts our travel from Tamil thesaurus to Tamil word net. It is a lexical resource which amalgamates all sorts of information available in a dictionary, thesaurus and wordNet. The Dravidian wordNets (in which Tamil wordNet is one of the four components) built under the IndoWordNet project depends on an ontology developed by western conceptualization of vocabularies of a language i.e. English). This has not taken into consideration the Indian conceptualization of vocabulary depicted in the *nikhandu* tradition. Say for examples, *nikhandus* have classifications such a six types of tastes, nine types of planets (*gragams*), 7 types of *mandalams* (a type of division), 15 *tidis* (15 phases of moon), etc which are crucial for Indian tradition. In the western oriented WordNet ontology there is no scope for the visualization of concepts depicted in *nikhandus*. More over building a wordNet based on Hindi wordNet which in turn is built on English wordNet will take many years to complete and it would miss the conceptualization depicted in Indian tradition. The present onto-thesaurus is based on the Indian conceptualization of vocabulary and the process of building one is very simple. We have the plan to make it into a generic one so that all the Dravidian languages can be easily accommodated into it.

1. Introduction

A paper thesaurus for Tamil was prepared in 1990 based on the principles of componential analysis of meaning propounded by Nida (1975) and was published in 2001 (Rajendran, 2001), nearly after a decade. Following the paper thesaurus, an Electronic thesaurus for Tamil was attempted and a book on Tamil electronic thesaurus was published in 2006 (Rajendran and Baskaran, 2006) The preparation of wordNet for Tamil was undertaken (2001-2003) with the financial assistance from Tamil Virtual University (renamed now as Tamil virtual academy) and a crude version of it based on the ontology developed by Rajendran (Rajendran, 2001) was submitted to the institute in 2003. After that, from 2009 onwards with the fund received from MHRD and Department of electronics and information Technology of Govt. of India the building of Dravidian wordNet was executed based on Hindi wordNet; nearly 3000 synsets (concepts) have been completed. Still we have to go a long way to achieve the desired target. At present a team from CEN, Amrita University is involved in building onto-thesaurus for Tamil as a part of the project entitled “Computing Tools for Tamil Language teaching and learning”. The project is funded by Tamil Virtual Academy, Chennai.

2. Tamil Onto-thesaurus

Thesaurus is a in its wider sense is a classification of words by concepts, topics, or subjects. The present Tamil Onto-thesaurus is the extended version of Electronic thesaurus of Tamil focusing more on the ontological features. Two kinds of issues arise in the preparation of Tamil onto-thesaurus:

- Linguistic issues
- Computational issues

2.1. Linguistic Issues

It involves mainly the following four tasks:

1. Developing an ontology for Tamil based on structural semantic principles.

2. Establishing semantic domains and sub domains based on distinguishing semantic or Componential features of lexical items.

3. Classifying Tamil vocabulary to fit into the ontology developed.

4. Linking words by various semantic or lexical relations such as synonymy, hyponymy-hyperonymy, meronymy-holonymy, compatibility, and incompatibiliity.

2.2. Computational Issues

It involves mainly the following three tasks:

1. Conversion of linguistic data base into computer accessible format.
2. Preparation of a tool to provide the facilities for augmenting, entering and editing the raw data, and classifying the lexical items in a semi-automatic way.
3. Creation of user friendly interfaces for accessing the onto-thesaurus in simple manner.

3. Ontology of Tamil vocabulary

The ontology available in Rajendran (2001), which is founded on the theory of componential analysis of meaning propounded by Nida (1975) is enhanced to suit the present purpose. The following is the skeletal structure of the Tamil ontology adopted in Onto-thesaurus.

1. பருப்பொருள்

எ. விலங்குகளல்லாதவை

1. இயற்கையானவை

எ. பூகோளம் சம்பந்தப்பட்டவை

பி. இயற்கைச் சாதனங்கள்

சி. தாவரங்கள் மற்றும் தாவரப்பொருட்கள்

2. உருவாக்கிய அல்லது கட்டப்பட்ட பருப்பொருள்கள்

எ. கலைப் பொருள்கள்

பி. பதன் செய்யப்பட்ட பொருள்கள்: உணவுகள் மற்றும் வாசனைப்பொருள்கள்

சி. கட்டிடங்கள்

பி. விலங்குப் பொருள்கள்

1. விலங்குகள், பறவைகள், பூச்சிகள்
2. மனிதர்கள்
3. இயற்கைக்கு அப்பாற்பட்ட சக்திகள் அல்லது உயிர்கள்

2. நிகழ்வுகள்

பௌதிகம் தொடர்பானவை, உடற்கூறு தொடர்பானவை, உணர்வு தொடர்பானவை, உணர்ச்சி தொடர்பானவை, அறிவு தொடர்பானவை, கருத்துப்பரிமாற்றம், சோ;க்கை, கட்டுப்பாடு, சலனம், தாக்கம், கைமாற்றம், தொடர்ச்சியான சலனங்கள் அல்லது செயல்கள் உள்ளடங்கிய பல்கூட்டச்செயல்கள்

3. அருவங்கள் காலம், தூரம், பருமை, விரைவு, சூடு, நிறம், எண், நிலை, மதப்பண்பு, கவர்ச்சி, வயது, உண்மை-பொய், நல்லது-கெட்டது, பலம், ஆரோக்கிய நிலை.

4. தொடர்பன்கள்

இடம் தொடர்பானவை, காலம் தொடர்பானவை, சுட்டல், தர்க்கம் தொடர்பானவை.

Nida (1975:15) advocates for four principal ways in which meanings of different semantic units can be related to one another. They are inclusion, overlapping, complementation and contiguity. These relations help us to classify the vocabulary of Tamil in terms of different semantic domains and distribute the lexical items in a structural fashion under these domains.

4. Structuring of vocabulary by lexical relations

Lexical semantics offers foundation for structuring vocabulary in terms of lexical relations (Lyons 230-335, Cruses 1986). In the NLP oriented papers, the general practice is to avoid giving linguistic details based on which the system is built. But here we would like to give the lexical semantics of building onto-thesaurus to make it more transparent.

4.1. Congruence Relations

The four basic relations between classes furnish a model for establishing fundamental group of sense relations and for defining a set of systematic variants applicable to virtually all other paradigmatic sense relations: Identity, inclusion, overlapping, and Disjunction (Cruse, 1986:86-87).

Identity: class A and class B have same members.

Inclusion: class B is wholly included in the class A

Overlap: class A and class B have members in common but each has members not found in the other

disjunction: class A and class B have no members in common

These four congruence relations culminates into the four lexical relations discussed below.

4.2. Lexical Relations

There are at least four lexical or meaning relations by which lexical items can be linked or related to one another in the ontological structure of Tamil vocabulary. They are synonymy, hyponymy, compatibility and incompatibility. A word acquires its referential meaning in being a member of a semantic domain by the common features it shares with other members in that domain, and having contrasting features which separate it from other members of the domain. It is the semantic relations among words, such as synonymy, hyponymy, compatibility and incompatibility (Cruse, 1986:84-111), which help one to classify and organize words in terms of semantic domains in a structural fashion.

Synonymy (ஒருபொருள்பன்மொழியம்): e.g. புத்தகம் 'book': *தூல்* 'book'; hyponymy-hypronymy உள்ளடங்குமொழியம்-உள்ளடக்குமொழியம்: e.g. *பசு* 'cow*': விலங்கு* 'animal'; meronymy-holonymy (சினைமொழியம்-முழுமொழியம்): கால்: உடல் ; Compatibility (இணக்கம்): e.g. செல்லப்பிராணி 'pet': நாய் 'dog'; Incompatibility (இணக்கமின்மை): e.g. மலை 'hill': மடு 'water hole'.

Incompatibility leads to the relation called opposition which culminates into many types which are discussed below.

4.3. Lexical Inheritance

Hyperonymy-hyponymy and meronymy-holonymy assure lexical item to inherit semantic features as exemplified below:

குயில் – இனிய குரலுடைய கரிய நிறப்பறவை.

பறவை -இரு கால்களும் அலகுகளும் உடைய, உடலில் இருபக்கங்களிலும் பறப்பதற்கு ஏற்றவகையில் சிறகுகளும் கொண்ட விலங்கினம்.

விலங்கினம் – தானாகவே இயங்கும் உணர்வு உறுப்புகளும் செல்லுலோஸ் இல்லாத செல் சுவர்களும் உள்ள உயிரினம்.

உயிரினம் – உயிர் வாழ்கின்ற ஒன்று.

4.4. Opposition

There are many types of oppositions (Lyons, 1977: 270-290, Cruse 197-263). They are listed below with examples.

Gradable opposites (நிரலாக்கம் செய்யவியலும் எதிர்மொழியம்**)**: e.g. நல்ல: கெட்ட; **Ungradable opposites (**நிரலாக்கம் செய்யவியலா எதிர்மொழியம்**)**: ஆண்: பெண்; **Complementaries** e.g. பகல்: இரவு, உண்மை: பொய், ஆண்:பெண், வெற்றி: தொல்வி; **Privative opposition (**தனிநிலை எதிர்மொழியம்): One denotes some positive property and the other denotes the absence of that property, e.g. உயிருள்ளவை]: உயிரில்லாதவை; **Equipollent opposition** (சமநிலை எதிமொழியம்): The contrasting lexemes denotes a positive propert e.g. ஆண்: பெண்; **Converseness** (மறுதலை)**:** Converseness is distinguished from antonymy and complementarity e.g, கணவன்: மனைவி, கொல்: கொல்லப்படு, வாங்கு: கொடு ; 'X Y-இன் கணவன்' என்பது 'Y X-இன் மனைவி' .

1. **Converse pairs of social roles:** e.g. : வைத்தியர்: நோயாள், எஜமானர்/எஜமானி: வேலையாள்; 2. **Converse pairs of Kinship terms** e.g**.** தந்தை/தாயார்:மகன்/மகள்; 3. **Converse pairs of Temporal relations:** e.g**.** முன்னால்: பின்னால், முன்னர்: பின்னர்; 4. **Converse pairs of Spatial relations**: e.g**.** முன்னால்: பின்னால், மேலே: கீழே

Directional opposites: e.g**.** மேலே : கீழே, வந்துசேர் 'arrive': புறப்படு; வா : போ; **Orthogonal opposites:** e.g**.** வடக்கு: கிழக்கு, மேற்கு; கிழக்கு: தெற்கு, வடக்கு; **Antipodal opposites:** வடக்கு: தெற்கு; கிழக்கு : மேற்கு

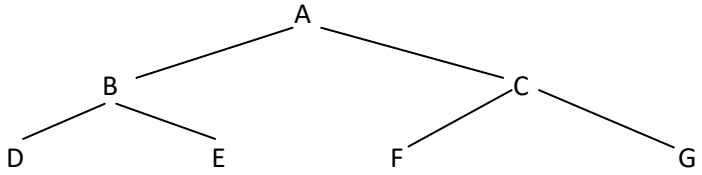
Non-binary oppositions

If the opposition involves more than two lexical items such contrasts are called non-binary opposition (Lyon, 1977:271). There a number of types of binary contrasts. They are dealt under non-branching hierarchies.

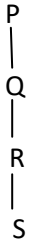
4.5. **Hierarchies**

Hierarchies are of two types: branching hierarchies and non-branching hierarchies. The branching hierarchy shows tree structure, where as non-branching hierarchy does not show tree structure (Cruse 1986:112).

Branching hierarchies

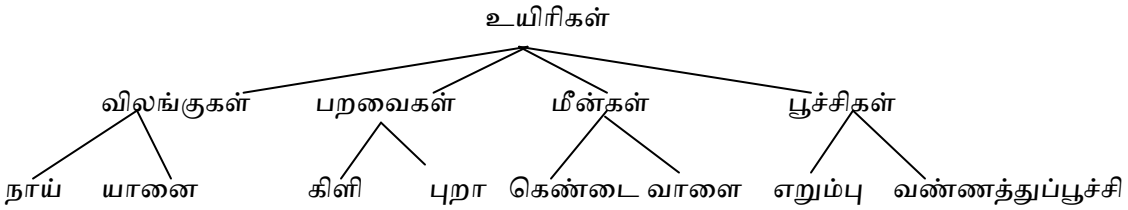


Non-branching hierarchies



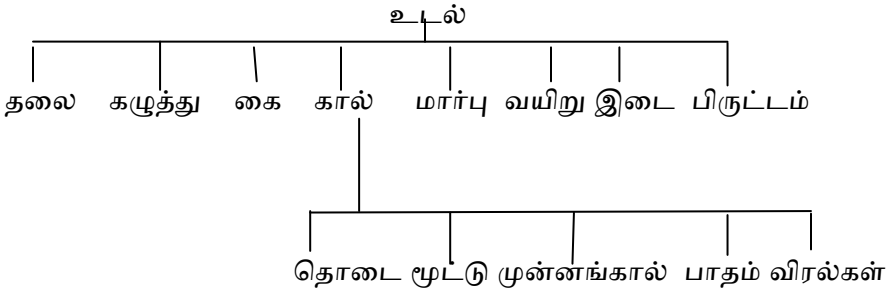
4.5.1. Taxonomic hierarchies

The hierarchies are the outcome of the hypronymy-hyponymy relation between lexical items. The sequence of hyperonymy-hyponymy relation leads the hierarchical structure of a set of vocabulary items which show this pair of relations among themselves. Taxonmic hierarchies are more liberal than hypronymy-hyponymy hierarchy. The following is an example:



4.5.2. Meronymic hierarchies

Meronymic hierarchies are the result of meronymy-holonymy relation shown by the lexical items. The meronymy-holonymy relation also gives hierarchical structure to a set of vocabulary. The following is an example.



4.5.3. Non-branching hierarchies

Non-binary opposition leads to a number of types of non-branching hierarchies. They are listed below:

Bipoles

The bipoles (Cruse, 2000: 189) is the simplest kind of linear structure found in a pair of opposites. They are simply oppositions which we have discussed earlier.

Bipolar chains

The bipolar chains (Cruse, 2000: 189) have implicit superlative terms of opposite polarity at each end of the scale. The following is the example: மிகநுண்ணிய, நுண்ணிய, சிறிய, பெரிய, மிகப்பெரிய

Monopolar Chains

In monopolar chains (Cruse , 2000: 190) there is no sense that the terms at the ends of the chain are oriented in opposite directions. Degree, stages, measures, ranks and sequences come under monopolar chains.

Degrees: Degrees (Cruse, 2000: 190) incorporate as part of their meaning different degrees of some continuously scaled property such as size or intensity. e.g. பாறை: குன்று: மலை: மாமலை

Stages: Stages (Cruse 2000: 190) are points in a life cycle of something and normally involve the notion of progression. e.g. பட்ட முன்படிப்பு, இளநிலைப் பட்டம், முதுநிலை பட்டம்;குழந்தைப் பருவம், வாலிபப் பருவம், முதிர் பருவம், முதுமைப் பருவம்; முட்டை, லார்வா, பியூபா, வண்ணத்துப்பூச்சி

Measures: Measures (Cruse, 2000: 190) are based on part-whole relationship, with each whole divided into a number of identical parts. e.g. செகண்டு, மினிட், மணி, நாள், வாரம், மாதம், ஆண்டு ; மில்லிமீட்டர், சென்டிமீட்டர், மீட்டர்; இஞ்சு, அடி, கஜம், பர்லாங்கு, மைல்; உழக்கு, ஆழாக்கு, பக்கா, மரக்கால்

Ranks: : The lexical items under ranks entails a sequential order which is not gradual (Cruse, 2000: 191): e.g. விரிவுரையாளர், முதுநிலை விரிவுரையாளர், இணைநிலை பேராசிரியர், பேராசிரியர்

Sequence: The lexical items which are ordered but does not bear increasing property are sequences (Cruse, 2000: 191). e.g. ஒன்று, இரண்டு, மூன்று, …, இறந்தகாலம், நிகழ்காலம், எதிர்காலம்; கார்காலம், கூதிர்காலம், முன்பனிக்காலம், பின்பனிக்காலம், இளவேனிற்காலம், முதுவேனிற்காலம்

Cyclical sets or cycles: The sequential lexical item can entail a cyclical order of time in the natural arena (Cruse 1986:187-190). e.g. {கார்காலம், கூதிர்காலம், முன்பனிக்காலம், பின்பனிக்காலம், இளவேனிற்காலம், முதுவேனிற்காலம்}; {சனவரி, பெப்ரவரி, மார்ச், ஏப்ரல், ஜூன்…}; {ஞாயிறு, திங்கள், செவ்வாய், புதன், வியாழன், வெள்ளி, சனி}

Propositional series or grids

The units of a gird is the cell, which consists of four lexical items, any one of which must be uniquely predictable form the remaining three (Cruse, 1986:118-131; Cruse, 2000: 191-193) . The followings are the examples of cells: e.g. பசு: கன்று :: குதிரை: குட்டி; பாம்பு: குஞ்சு :: பல்லி: குஞ்சு; யானை : பிளிறு :: நரி: உளையிடு

5.User friendly interface for accessing the onto-thesaurus

A few user friendly interfacs have been prepared from which one access the infomration needed form the onto-thesaurus of Tamil.

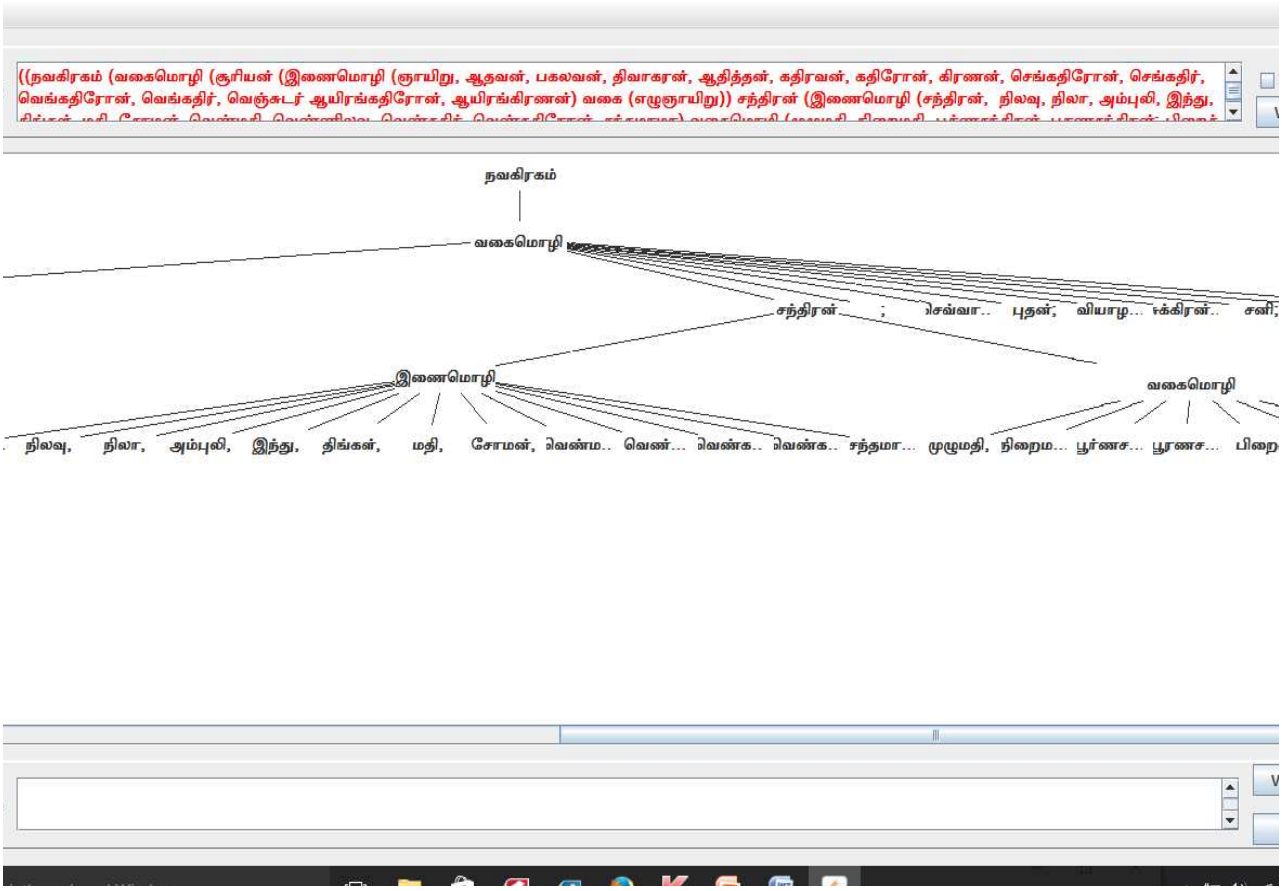
5.1. Linguistic Tree Viewer in Java

NLP or Linguistic researchers who work in syntax often want to visualize parse trees or create linguistic trees for analyzing the structure of a language. TreeViewer software provides an easy to use interface to visualize or create simple linguistic trees. This software is written entirely in Java. This tree viewer has been converted to depict the ontological structure of Tamil vocabulary. The semantic relations such as synonymy, hyponymy-hypernymy, meronymy-holonymy, oppositions, entailments are captured by the tree viewer. The tree representation will be converted into an ontology based visual thesaurus.

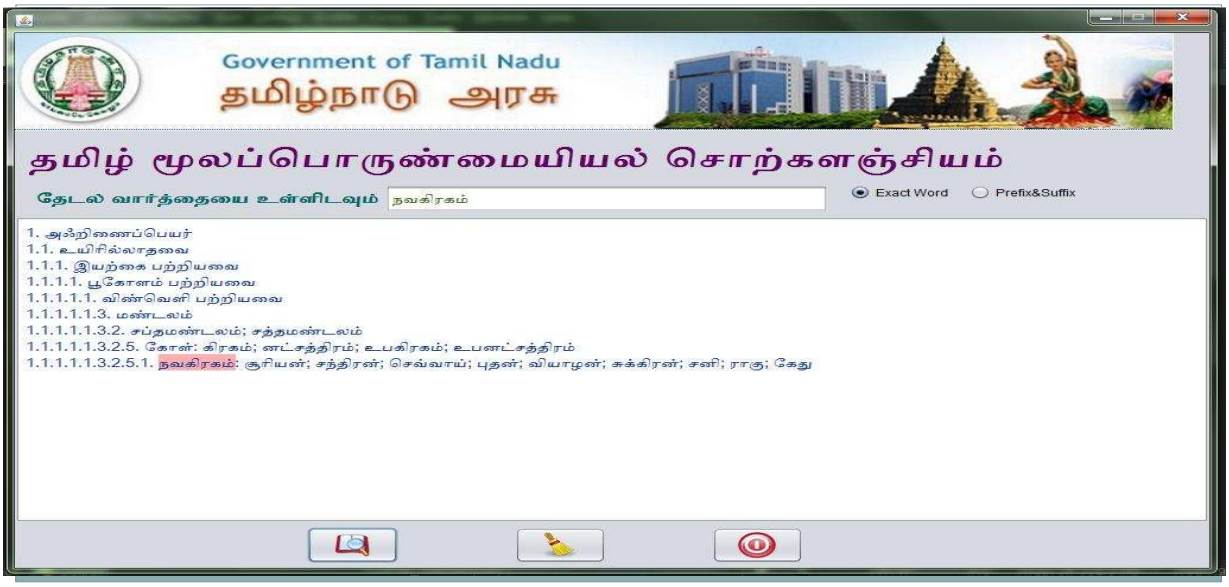
5.2 Tree structure of a domain

As a sample for demo the semantic domain *nava kirangkaL* 'nine planets' has been taken and converted into tree structure as shown in the snap-shot given below:

((நவகிரகம் (வகைமொழி (சூரியன் (இணைமொழி (ஞாயிறு, ஆதவன், பகலவன், திவாகரன், ஆதித்தன், கதிரவன், கதிரோன், கிரணன், செங்கதிரோன், செங்கதிர், வெங்கதிரோன், வெங்கதிர், வெஞ்சுடர் ஆயிரங்கதிரோன், ஆயிரங்கிரணன்) வகை (எழுஞாயிறு)) சந்திரன் (இணைமொழி (சந்திரன், நிலவு, நிலா, அம்புலி, இந்து, திங்கள், மதி, சோமன், வெண்மதி, வெண்ணிலவு, வெண்கதிர், வெண்கதிரோன், சந்தமாமா) வகைமொழி (முழுமதி, நிறைமதி, பூர்ணசந்திரன், பூரணசந்திரன்; பிறைச் சந்திரன், பிறை)); செவ்வாய்; புதன்; வியாழன்; சுக்கிரன்; சனி; ராகு; கேது)))))



There is one more GUI which gives the hierarchical details of a lexical items there by one will be able to get the meaning of the lexical item form the hierarchical information. This shown in the following screen shot.



6. Conclusion

The coverage of the vocabulary at present is only 30000 lexical items. We hope to improve on it in the near future. We like to accommodate all kinds of lexical and meaing relations or linkages a user expects from the Tamil onto-thesaurs. All the information availabel to a word and a set of words will be incorporated in the onto-thesarus of Tamil. The present onto-thesurus system will be converted into a generic system so as to accommodate all the other Dravidian languaes. The onto-thesaurus under preparation has wide range of uses which inclue information retrieval across Dravidian languages, machine translation across dravidian languages and building knowledge systems for Dravidian languages.

References

(1) இராசேந்திரன். ச. 2001. தற்காலத் தமிழ்ச் சொற்களஞ்சியம். தஞ்சாவூர்: தமிழ்ப் பல்கலைக்கழகம்.

(2) இராசேந்திரன், ச. & பாஸ்கரன், ச. தமிழ் மின்சொற்களஞ்சியம். 2006. தஞ்சாவூர்: தமிழ்ப் பல்கலைக்கழகம்.

(3) Nida, E.A. 1975. Componential Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mouton.

(4) Cruse, D.A. 1986. Lexical semantics. Cambridge: Cambridge University Press.

(5) Cruse, D. A. Meaning in Language: An Introduction to semantics and pragmatics. Oxford: Oxford University Press.

Lyons, J. 1977. Semantics, volume 1, Cambridge: Cambridge University Press.